Liu, C.-C. et al. (Eds.) (2014). Proceedings of the 22nd International Conference on Computers in Education. Japan: Asia-Pacific Society for Computers in Education

A Tagging Editor for Learner Corpora Annotation and Error Analysis

Lung-Hao LEE [a,d], Kuei-Ching LEE [a,d], Li-Ping CHANG [b], Yuen-Hsien TSENG[a*], Liang-Chih YU[c,d] & Hsin-Hsi CHEN[e]
[a]*Information Technology Center, National Taiwan Normal University, Taiwan*
[b]*Mandarin Training Center, National Taiwan Normal University, Taiwan*
[c]*Department of Information Management, Yuen-Ze University, Taiwan*
[d]*Innovation Center for Big Data and Digital Convergence, Yuen-Ze University, Taiwan*
[e]*Department of Computer Science and Information Engineering, National Taiwan University, Taiwan*
*samtseng@ntnu.edu.tw

Abstract: In this paper, we describe the development of the tagging editor for learner corpora annotation and computer-aided error analysis. We collect essays written by learners of Chinese as a foreign language for grammatical error annotation and correction. Our tagging editor is effective and enables the annotated corpus to be used in a shared task in ICCE 2014.

Keywords: Computer-aided error analysis, learner corpora, interlanguage, Mandarin Chinese

1. Introduction

Learner corpora are the collection of foreign language learners' produced responses, which are valuable resources for research of second language learning and teaching. For example, the International Corpus of Learner English (ICLE) is considered as one of the most important learner corpora. ICLE consists of argumentative essays written by advanced learners of English as a Foreign Language from different native language backgrounds (Granger, 2003). The first version was published in 2002. They are currently working towards the third version of this corpus. In addition, the Cambridge Learner Corpus (CLC) is made up of more than 200 thousands of examination scripts written by English learners speaking 148 different mother languages (Nicholls, 2003). CLC was established to assist English Language Teaching/Training (ELT) publishers to create aided materials, e.g., Cambridge dictionaries and ELT course books, addressed to their target users.

Annotated learner corpora play an important role to develop natural language processing techniques for educational applications. As an example, Assessing LExical Knowledge (ALEK) system (Chodorow and Leacock, 2000) adopted statistical analysis from learner corpora to detect the errors of an English sentence. Izumi et al. (2003) detected English grammatical and lexical errors made by Japanese learners. Lee et al. (2013) proposed a linguistic rule based approach to detect grammatical errors written by learners of Chinese as a foreign language. Recently, the CoNLL 2013/2014 shared tasks focus on grammatical error correction for learners' English as a foreign language (Ng. et al. 2013). SIGHAN 2013/2014 bakeoffs on Chinese spelling check evaluation focus on developing automatic checker to detect and correct spelling errors (Wu et al. 2013).

However, to make learner corpora useful for these tasks, they must be annotated correctly before automated analysis can be applied. In this work, we design a tagging editor to help annotators to insert error tags and rewrite correct usages for the sentences in the learner corpus. In addition, our editor provides the function for error analysis, which further assists annotators to instantly discover incorrect or inconsistent tagging instances during the annotation process.

2. The Error Tagging Editor

The construction of a tagging editor includes designing tag-associated error categories arranged on a menu interface, which can help annotators to select and insert error tags alongside the wrong part of

the learners' written texts. In addition to error tagging, reconstruction of correct usages is usually needed in the annotation process. After the learner corpus is tagged and corrected, error analysis can be done quantitatively according to various kinds of users' interests.

Figure 1 shows a screenshot of the tagging editor. The functions of our tagging editor can be divided into three main parts: (1) Searching zone (the left panel): learners' written texts are stored in individual files accompanying with their metadata, such as the level describing the learner's language proficiency, the score of the learner's written texts, and the learner's mother-tongue language (ML). Our tagging editor can search learners' texts using these fields of metadata. The searching results can be listed in order by the unique ID with the (tagging | correction) status. The symbol "〇" means a finished situation. In contrast, the symbol "X" represents that the texts need to be annotated. (2) Tagging zone (the middle panel): when the texts are loaded in this zone, annotators can select and insert error tags from the menu bar into some position of learners' texts. Inserted tags are shown in terms of square brackets in red color. (3) Correction zone (the right panel): annotators usually need to correct the error parts for correct usages. Correction zone is aligned paragraph-wisely with tagging zone to facilitate annotators' corrections. We highlight the changed texts in blue color. Besides, our tagging editor also reports error analysis, which benefits annotators to find incorrect/inconsistent tagging instances to be fixed in the verification procedure.

Our tagging editor is flexible enough to meet various annotation tasks for learner corpora in different language. The character encoding is in Unicode; the editor is developed in Java; both of which are cross-platform. Besides, annotators can add, delete or fix their self-defined error tags for their annotation tasks. The metadata is also optional. The tagging editor could load the learners' written texts even without the accompanying metadata.

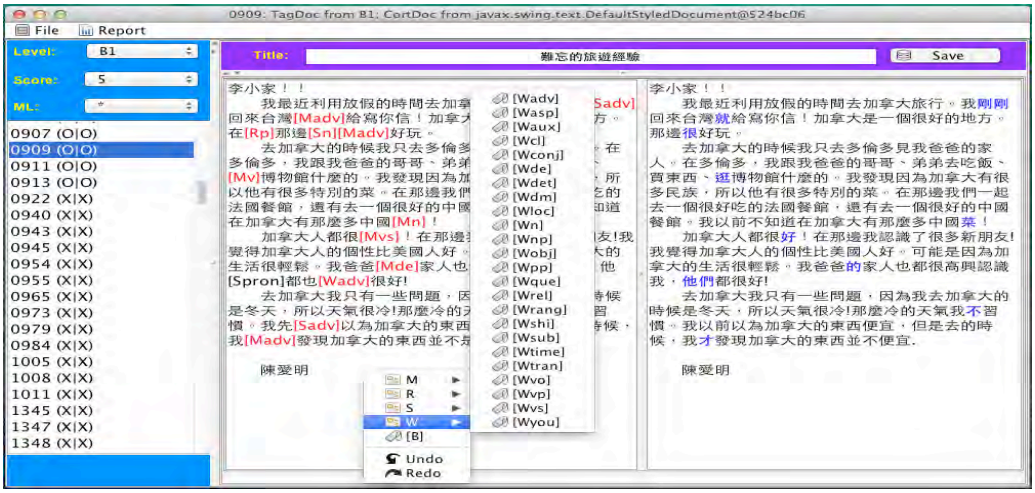

Figure 1. Screenshot of our tagging editor

3. The Annotation of TOCFL Learner Corpus

The annotated corpus using this tagger editor is mainly from the computer-based writing Test of Chinese as a Foreign Language (TOCFL). The writing test is designed according to the six proficiency levels of the Common European Framework of Reference (CEFR). Test takers have to complete two different tasks for each level. For example, for the A2 (Waystage level) candidates, they will be asked to write a note and describe a story after looking at four pictures. All candidates are asked to complete the writings on line. Each text is then scored on a 0-5 point scale. Score 5 means high-quality writings, score 3 is the threshold for passing the test, and so forth. There are 4,567 essays collected in the corpus so far.

For the purpose of studies in Chinese learners' interlanguage, hierarchical error tags are designed to annotate grammatical errors. There are two types of error tagging, one is target modification taxonomy, the other is linguistic category classification. The former includes four PADS error types: mis-ordering (Permutation), redundancy (Addition), omission (Deletion), and mis-selection (Substitution). The latter includes 36 linguistic types, e.g., noun, verb, preposition,

specific construction, and so on. Using our tagging editor, 51 essays belonging to CEFR B1 level with score 5 had been annotated by two linguists at the same time. Figure 2 shows the distribution of error tags sorted by their occurrence. In total, there are 678 errors in 51 essays. The top 3 error tags are: Sv (mis-Selection of verbs), Madv (Missing of adverbs), and Sn (mis-Selection of nouns). Their frequency is 53, 40, and 39, respectively. In the above error tag abbreviations, the first capital letter of the tag represents the higher level of error tagging.

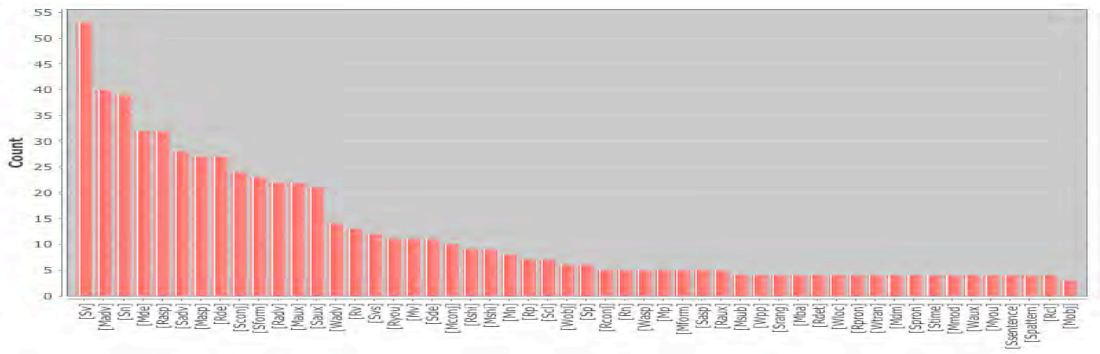

Figure 2. Distribution of error tags in the annotated corpus

4. Conclusions and Future Work

This article describes our tagging editor that can be employed to meet annotation tasks in learner corpora. This tagging editor is effective to annotate grammatical errors in CFL learners' essays. We will further collect annotators' feedbacks and discuss with them to enhance its functions. We shall release this editor for public use in the future. The corpus annotated by this editor described above has currently been used for a shared task in the ICCE workshop.

Acknowledgements

This research was supported by the Ministry of Science and Technology, under the grant MOST 102-2221-E-002-103-MY3, 102-2221-E-155-029-MY3, 103-2221-E-003-013-MY3, 103-2911-I-003-301 and the "Aim for the Top University Project" and "Center of Learning Technology for Chinese" of National Taiwan Normal University, sponsored by the Ministry of Education, Taiwan, R.O.C..

References

Chodorow, M., & Leacock, C. (2000). An unsupervised method for detecting grammatical errors. *Proceedings of NAACL'00* (pp. 140-147). Seattle, Washington: ACL Anthology.

Díaz-Negrillo, A., & Fernández-Domínguez, J. (2006). Error tagging systems for learner corpora. *RESLA*, 19, 83-102.

Granger, S. (2003). The International Corpus of Learner English: a new resource for foreign language learning and teaching and second language acquisition research. *TESOL Quarterly*, 37(3), 538-546.

Izumi, E., Uchimoto, K., Saiga, T., Supnithi, T., & Isahara, H. (2003). Automatic error detection in the Japanese learner's English spoken data. *Proceedings of ACL'03* (pp. 145-148). Sapporo, Japan: ACL Anthology.

Lee, L.-H., Chang, L.-P., Lee, K.-C., Tseng, Y.-H., & Chen, H.-H. (2013). Linguistic rules based Chinese error detection for second language learning. *Proceedings of ICCE'13* (pp. 27-29). Bali, Indonesia: Asia-Pacific Society for Computers in Education.

Ng, H. T., Wu, S. M., Wu, Y., Hadiwinoto, C., & Tetreault, J. (2013). The CoNLL-2013 shared task on grammatical error correction. *Proceedings of CoNLL'13* (pp. 1-12), Sofia, Bulgaria: ACL Anthology.

Nicholls, D. (2003). The Cambridge Learner Corpus – error coding and analysis for lexicography and ELT. *Proceedings of CL'03* (pp. 572-581), Lancaster, UK.

Wu, S.-H., Liu, C.-L., & Lee, L.-H. (2013). Chinese spelling check evaluation at SIGHAN bake-off 2013. *Proceedings of SIGHAN'13* (pp. 35-42), Nagoya, Japan: ACL Anthology.